Whole genome sequencing from the New Zealand *Saccharomyces cerevisiae* population reveals the genomic impacts of novel microbial range expansion

Peter Higgins[1], Cooper A Grace[2,3], Soon A. Lee[4], Matthew R Goddard[1,4]

1 The School of Life Sciences, University of Lincoln, Lincoln, United Kingdom

2 Department of Biology, York Biomedical Research Institute, University of York, Heslington, York, United Kingdom

3 Department of Biological and Geographical Sciences, University of Huddersfield, Huddersfield, United Kingdom

4 The School of Biological Sciences, University of Auckland, Auckland, New Zealand

Abstract

Saccharomyces cerevisiae is extensively utilised for commercial fermentation, and is also an important biological model; however, its ecology has only recently begun to be understood. Through the use of whole genome sequencing, the species has been characterised into a number of distinct subpopulations, defined by geographical ranges and industrial uses. Here the whole genome sequences of 104 New Zealand *S. cerevisiae* strains, including 52 novel genomes, are analysed alongside 450 published sequences derived from various global locations. The impact of *S. cerevisiae* novel range expansion into New Zealand was investigated and these analyses reveal the positioning of New Zealand strains as a sub-group to the predominantly European/wine clade. A number of genomic differences with the European group correlate with range expansion into New Zealand, including 18 highly enriched SNPs and novel Ty1/2 insertions. While it is not possible to categorically determine if any genetic differences are due to stochastic process or the operations of natural selection, we suggest that the observation of New Zealand specific copy number increases of four sugar transporter genes in the *HXT* family may reasonably represent an adaptation in the New Zealand

S. cerevisiae subpopulation, and this correlates with observations of copy number changes during adaptation in small scale experimental evolution studies.

Introduction

Most plant and animal species have defined geographic ranges which they inhabit. The invasion of ecologically similar but geographically different habitats, to which species are reasonably well preadapted, may occur if migration and the absence of competition allows. The invasion of new ranges by species can have major impacts on genetic diversity and population structure (Braga et al., 2019), including admixture, adaptive evolution, and neutral evolution via genetic drift and bottleneck events (Keller et al., 2010; Swaegers et al., 2015; e.g. Garcia-Elfring et al., 2017). The population genetic effects of range expansion have been reported in both plant species, e.g. accumulation of deleterious mutations in *Mercurialis annua* (González-Martínez et al., 2017), and animal species, e.g. poleward expansion in the damselfly *Coenagrion scitulum* (Swaegers et al., 2015). However, the concept of range expansion in microbes has only been more recently explored, and its impact on microbial population structure and evolution is not as well understood. Most studies have focused on pathogenic species – primarily plant-pathogenic fungi (for review, see Thakur et al., 2019). Like many other aspects of microbial ecology, range expansion of microbes is poorly characterised due to their cryptic nature (Jarić et al., 2019). It is now clear that Baas-Becking's classic microbial biogeographic hypothesis of '*Everything is everywhere:* but *the environment selects*' does not hold universally for microbes (van der Gast, 2015). While some microbes appear to be dispersed ubiquitously, others have restricted distributions, as described by a moderate endemicity model of microbial biogeography (Foissner, 2007). This suggests that, as with macroscopic species, geographic separation has the potential to drive population structure alongside environmental processes (O'Malley, 2008). However, the factors that constrain or define microbial range are not always clear, and the relative contribution of ecological and geographical factors is still being evaluated (Power et al., 2018; Meyer et al., 2018).

Saccharomyces are unicellular diploid eukaryote fungi which undergo sexual reproduction, and these yeasts have a long historical association with human populations due to their ability to produce CO_2 and ethanol via sugar fermentation. *Saccharomyces cerevisiae* has been involved in the production of fermented beverages for at least 9000 years (McGovern et al., 2004). The global phylogeny of *S. cerevisiae*, based on partial and whole genome sequencing, shows it to be grouped both by technological origins (including wine, beer, bread, cheese, rum, and sake), and geographic area (European, North American Oak, West Africa, Malaysia, and the Far East) (Legras et al., 2007; Liti et al., 2009; Wang et al., 2012; Peter et al., 2018). The general picture is that *S. cerevisiae* originated in Far East Asia (Wang et al., 2012; Duan et al., 2018) and has expanded and diversified to now comprise specific geographic sub-populations. However, it is not clear what forces maintained these geographic sub-populations. In addition, several distinct lineages have been inadvertently domesticated by humans for food and beverage fermentations from these subpopulations (Duan et al., 2018; Steensels et al., 2019). Among the most studied of these are the *S. cerevisiae* strains associated with viticulture and winemaking. Recent work has shown this group to be domesticated from a wild population inhabiting the Mediterranean region, presumably alongside the development and expansion of viticulture there (Almeida et al., 2015). As viticulture has since spread throughout the world, this group of vine and wine associated yeasts (the wine-group) were brought with it.

Population structure and genetic diversity of the *S. cerevisiae* wine-group has been associated with geographic location at global and regional scales (Martínez et al., 2007; Viel et al., 2017), grape variety (Schuller et al., 2012), and environmental conditions, e.g. freeze-thaw resistance and copper sulfate resistance (Kvitek et al., 2008). Research has indicated the diversity of the wine-group originates from dispersal and expansion into new environments due to the novel selection pressures encountered, such as metabolic stress in industrial environments (Dunn et al., 2012). Adaptive variation may include gene gain or loss (Almeida et al., 2017), mutations within genes (Aa et al., 2006), as well as copy

number variation of genes to mediate regulation of gene products (Dunn et al., 2012). However, it has also been suggested that the genetic differences observed between *S. cerevisiae* populations may be due to neutral processes, primarily population bottlenecks and genetic drift (Warringer et al., 2011). This view has been supported by a number of studies (Liti et al., 2009; Zörgö et al., 2012) and forms the basis of the neutral nomad model of *S. cerevisiae* (Goddard and Greig, 2015). Here the genetic diversity of *S. cerevisiae* is explained as neutral divergence through isolation by distance, as observed by the lack of gene flow between geographically remote populations. It seems highly probable that both neutral and adaptive processes have contributed to the genetic architecture of modern *S. cerevisiae* lineages, but establishing their relative contributions is difficult without further understanding the life history of these lineages. Before making ecological inferences about genomic features, it is vital to ascertain whether specific genomic features have arisen through adaptive or neutral processes.

As one of the last landmasses colonised by humans, ~700 years ago (King, 2003), and one of the most geographically isolated, New Zealand (NZ) provides a unique opportunity to study the effects of range expansion for species generally and is thus also ideal to evaluate microbial range expansion. Extensive sampling in New Zealand vineyards, wineries and native forests have discovered approximately 2,000 genotypes of *S. cerevisiae* comprising regionally distinct sub-populations at scales greater than 1,000km, inferred using variance at micro-satellite loci (Goddard et al., 2010; Knight et al., 2015). Microsatellite and Rad-Tag DNA sequence (Cromie et al., 2013) analyses suggest that the NZ population is closely related to, but distinct from, the wine-group, whereas analyses of 106 conserved loci derived from 50 whole genome sequences of these isolates infers these to be mostly interdigitated with the wine-group, and this group expanded to New Zealand via human aided dispersal within only the last 1,000 years (Gayevskiy et al., 2016). The status of the NZ population is thus not clear.

Transposable elements (TEs) are mobile components of almost all eukaryotic genomes, and their evolutionary history typically follows that of their hosts (Bowen et al., 2003; Hosid et al., 2012). *S. cerevisiae* contains multiple families of the TE subclass known as long-terminal repeat (LTR) retrotransposons, named *Ty1-5* (Kim et al., 1998), that drive their own replication and integration into the genome via an RNA intermediate. LTR sequences frame *gag* and *pol* open reading frames (ORFs) which encode proteins necessary for their transposition (reviewed by Havecker et al., 2004). LTR sequences are identical upon transposition, and as such present a target for homologous recombination, which results in the deletion of the ORFs and one LTR, leaving a 'solo' LTR remnant in the genome (Lesage and Todeschini, 2005). Hybridisation between the closely related elements of the *Ty1-2* superfamily is also thought to occur via recombination (Jordan and McDonald, 1999; Czaja et al., 2020). Analysis of the complex familial histories by Carr et al. (2012) revealed that elements of the *Ty1-2* superfamily are still highly active, members of *Ty3* and *Ty4* transpose at a relatively low rate and *Ty5* is extinct in *S. cerevisiae*. Insertion sites of these elements have therefore been used to infer evolutionary relationships in *S. cerevisiae,* with insertions shared between populations providing strong evidence for common ancestry and horizontal transfer from other species providing evidence of geographical background and life histories (Carr et al., 2012). Bleykasten-Grosshans et al. (2013) analysed the *Ty* insertion profiles of geographically and ecologically diverse strains of *S. cerevisiae to* identify the patterns of *Ty* distribution which contribute to their diversity. However, little investigation into the evolutionary history of *Ty* elements in an isolated population such as NZ has been conducted.

In this study, an additional 52 *S. cerevisiae* genomes are sequenced from the NZ population, and combined with the existing NZ derived data to total 104 high-quality genomes. Here the entirety of genomes (as opposed to select loci used in previous studies (Gayevskiy et al., 2016)) are used to evaluate the micro-evolutionary nature of microbial range expansion into NZ. Signals of differentiation are examined to determine the effects of range expansion on *S. cerevisiae* genetic diversity and

population structure. The origins of these signals are investigated, and we attempt to infer if these are from natural selection, or if divergence within NZ is primarily due to neutral processes. These analyses will also allow the placement of NZ *S. cerevisiae* within the wine clade to be more accurately estimated.

Methods

Strain selection

The sampling locations of the 104 strains were across major wine growing areas on NZ's north and south islands (Table S1), and all but one strain derived from spontaneous wine ferments or vineyard habitats, the one exception was isolated from native forest soil. 52 unsequenced strains were selected to complement those genomes analysed by Gayevskiy et al. 2016 (originally reported in Goddard et al., 2010; Knight and Goddard, 2015). Each strain was propagated in YPD, and high molecular weight genomic DNA was extracted using the Qiagen Blood & Cell Culture DNA Kit. Libraries were constructed using the Illumina TruSeq Nano DNA Sample Prep Kit with 550 bp insert size. Sequencing was carried out at the Beijing Genomics Institute (China) on three 150-bp paired-end lanes of an Illumina HiSeq 2000.

Genome sequencing and read alignment

FASTQC v0.11.7 (Andrews, 2010) was used to analyse the quality of libraries and identify trimming parameters. Reads were trimmed using Trimmomatic v0.36 (Bolger et al., 2014) with parameters LEADING:3 TRAILING:3 SLIDINGWINDOW:3:20 MINLEN:30 and ILLUMINACLIP to remove adapter sequences. Resulting trimmed reads were compared to raw reads to confirm quality improvement. Trimmed reads were aligned to the reference strain S288C using Bowtie2 v2.3.4.1 (Langmead and Salzberg, 2012). Aligned Bam files were sorted and aligned using Samtools v1.8 (Li et al., 2009).

Duplicate reads were removed with Samtools markdup function. Variants were called with bcftools mpileup with default parameters, and bcftools call with parameters -mv --ploidy 2. The Samtools package bcftools filter was used to remove reads with quality scores less than 20.

Population structure analysis

Prior to structural analysis, samples were filtered to reduce computational time and remove uninformative SNPs. Using vcftools v0.1.16 (Danecek et al., 2011) with flags --maf 0.02 --max-maf 0.98 --max-missing-count 0 --thin 50, rare SNPs, SNP sites missing data in any sample and any SNPs within 50bp of each other were removed. PLINK v1.90b4 (Purcell et al., 2007) variant pruning was used with --indep-pairwise 50 5 0.9 and –hwe 1e-10 to filter out SNPs with high Linkage Disequilibrium (LD) and those far from Hardy–Weinberg equilibrium. To identify population structure within NZ, Structure v2.3.4 (Pritchard et al., 2000) was run using both Admixture and Linkage models on all 104 strains. Values for the number of assumed clusters (K) were tested from 2 to 12, with 5 repeat runs at each value of K. Population classifications were not used for the prior. Iteration numbers were increased until convergence was achieved for all values of K. The resulting Structure data were analysed in CLUMPAK v1.1.2 (Kopelman et al., 2015) to test for group modality and aligning results over all values of K, and the Evanno method was employed to determine the optimal value of K (Evanno et al., 2005). To test for biogeographical structure, Structure output for the optimal K value was analysed in ObStruct v1.0 using either sampling locations or North/South Island as predefined populations (Gayevskiy et al., 2014). An additional Structure analysis was conducted using the Admixture model on 15272 SNPs from a randomised selection of 50 NZ strains, 40 EU strains – including 5 from each defined subpopulation – and 5 from both alpechin and mosaic populations, for K values of 1 to 4.

Analysis of variant SNPs

SNP data for published strains was taken from Peter et al. (2018) with SNPs extracted from the provided gvcf file. Data for 362 wine-group strains were extracted and combined with NZ vcf files before conversion to gds format using gdsfmt v1.22.0 (Zheng et al., 2012). The combined dataset was filtered to remove SNPs with missing rates greater than 0.9 and with linkage disequilibrium (LD) greater than 0.5. Additionally, the NZ data was filtered similarly for use in independent analysis. Principal Component Analysis (PCA) was performed on both the NZ data and the combined dataset using snpgdsPCA in SNPRelate v1.20.1 (Zheng et al., 2012). Samples were assigned to populations based on Structure results for K=2, 3 and 4. A sample was assigned to a population if it had greater than 0.8 inferred ancestry from that population. Samples with less than 0.8 inferred ancestry from any population were designated as admixed. Using the SNP loadings from this PCA, the wine-group strains were projected onto the NZ Principal Components with functions snpgdsPCASNPLoading and snpgdsPCASampLoading. Additionally, PCA was performed on the combined wine-NZ dataset. Two phylogenies were generated, the first included all strains from Peter et al. (2018) and all 104 NZ samples, the second included only NZ, wine-group, and closely related mosaic and alpechin (olive mill wastewater) strains. Phylogenies were created using hierarchical clustering in SNPRelate and visualised with iTOL v4 (Letunic and Bork, 2019).

Signals of SNP selection

To model differences in SNP frequency, BayeScan v2.0 was used on a subset of SNPs from both NZ and wine-group strains, filtered with vcftools to remove rare SNPs (those with minor or major allele frequencies below 10%) (Foll and Gaggiotti, 2008). BayeScan estimates the relative posterior probabilities for each locus being under either diversifying selection or balancing/purifying selection. Multiple runs of BayeScan were conducted with varying model parameters and the results of each run

were analysed to test for convergence, as confirmed with Gelman and Rubin's diagnostic in R package coda v0.19.3 (Plummer et al., 2006).

Transposable elements analysis

104 NZ genomes were screened with RepeatMasker (Smit et al, 2013) for the presence of *Ty1-5* LTRs using a custom library (see File S5). The genomic co-ordinates generated by RepeatMasker were converted to .bed format and used to extract full-length LTR sequences from contigs of each strain using Bedtools getfasta (Quinlan and Hall, 2010). Coordinates were modified in R (R Core Team, 2013) to include 5bp up/downstream to obtain target site duplications (TSDs) present at the 5' and 3' sequence boundaries. Partial LTRs and those whose TSDs could not be determined, i.e. due to the presence of Ns or placement at the end of a contig were discounted from analysis. NZ LTRs were sorted by superfamily and added to modified superfamilial LTR datasets compiled by Carr et al. (2012). Carr et al. (2012)'s datasets were obtained by screening the Sanger Genome Resequencing Project (SGRP) *S. cerevisae* and *S. paradoxus* genomes (available at https://www.sanger.ac.uk/research/projects/genomeinformatics/sgrp.html). The publicly available genomes of *S. kudriavzevii* (assembly AACI03), *S. eubayanus* (assembly TUD_Seub_ont_v2), *S. uvarum* (assembly AACA01; formerly *S. bayanus*) and *S. mikatae* (assembly AACH01) were also screened for LTR sequences. LTRs were aligned with MAFFT v7.310 (Katoh and Standley, 2013) using the FFT-NS-2 alignment method and –leavegappyregion as an additional parameter. Alignments were trimmed with trimAL v.1.2 (Capella-Gutiérrez et al., 2009) with the options -gt 0.9 and -cons 90 (with the exception of the *Ty1-2* alignment which used the option -cons 70). Identical sequences and LTRs shared between multiple strains (as determined by TSDs) were removed, with preference given to the SGRP strains. Alignments for each superfamily were converted to PHYLIP format for phylogenetic analysis with RaxML-HPC2 (Stamatakis, 2014) on XSEDE as part of the CIPRES Science Gateway (Miller et al., 2010). The following parameters were used: GTRGAMMA nucleotide bootstrapping model; 1000 bootstrap

iterations; random seed values for parsimony and rapid bootstrapping. All other parameters were kept as default. Resulting trees were visualised with FigTree v. 1.4.4 (Rambaut, 2012).

Analysis of Copy Number Variants

To search for signatures of selection, NZ genomes were mapped to the published *S. cerevisiae* ORF pangenome using bwa v0.7.17 with the -U 0 flag set, so as to not penalise unpaired reads due to the shorter targets (Li and Durbin, 2009). This ORF pangenome was constructed using de novo assembly of 1,011 *S. cerevisiae* genomes and consists of 7,796 ORFs, with closely related and duplicate sequences removed (Peter et al., 2018). The copy number of each ORF was determined by dividing the median read depth at each ORF by the median read depth for that sample. These results were compared to the ORF copy numbers given for 362 strains in the wine-group. Median copy numbers were calculated for each population. ORFs were filtered to include only those with a median copy number of at least 0.5 in either population. Those ORFS enriched in NZ by a factor of 2 or greater were selected for Gene Ontology enrichment analysis. GOrilla was used to search for enriched Process, Function and Component GO terms against a background of all verified *S. cerevisiae* genes (Eden et al., 2009). The presence/absence of the 2µ plasmid and its relative copy number was calculated as the median of plasmid ORF copy numbers.

Data availability Statement

Raw sequence data has been made publicly available at the NCBI Sequence Read Archive with accession PRJNA649809. Supplementary material is available at Figshare. Figure S1 shows hierarchical clustering analysis for 1115 *S. cerevisiae* strains. Figure S2 shows Evanno method results for determining K. Figure S3-4 show phylogenies for Ty3 and Ty4. Files S1-4 are tree files for the Ty family.

File S5 is a custom Ty library. Table S1 shows sampling information. Table S2 includes a list of enriched ORFs in NZ. Table S3 shows additional Gene Ontology details.

Results

In total, 346 million 150-bp paired-end reads were obtained, giving an average of 6.7 million reads per genome. Of these, an average of 95.3% mapped to the S288C reference genome, with an average read depth of 63.9X over all covered regions and are thus comparable to those genomes generated by Gayevskiy et al. 2016. When combined with the data from Gayevskiy et al., 2016, 604 million mapped reads underlie the 104 genomes.

Global positioning of NZ isolates

An average of 44,960 homozygous and 6,540 heterozygous SNPs were identified across the 104 NZ genomes when compared to the S288C reference, with a transition/transversion ratio of 2.6. When compared with the 1,011 high quality genomes from Peter et al. 2018, the position of NZ strains in the wider wine clade and global *S. cerevisiae* phylogeny is shown in Figure 1 and Figure S1. Mosaic strains closely related to the wine-group and characterized by admixture from other lineages (Peter et al., 2018) were also included in the wine-group analysis. NZ strains do not appear randomised among the wine-group (green) but have a preponderance to cluster toward and with the mosaic clade at the root of the tree. To further elucidate the relationship of NZ strains, a PCA was performed on a subset of 16,672 SNPs, filtered to remove those in high linkage disequilibrium (LD) (Figure 2). The first two principal components explained 2.86% and 2.34% of the variance and clearly separate the vast majority NZ and wine-group strains. PC1 primarily distinguished a number of apparent wine-group subpopulations, and separated a single NZ sample, WTETETsf_C10, from the NZ group, possibly indicating the occurrence of more recent arrival (as has been shown previously by Goddard et al,

(2010) via oak barrels from Europe) or admixture. PC2 distinguished the NZ group from all but eight of the wine-group samples. Further principal components were visualized but were less informative, with PC3 and PC4 explaining only 2.1% and 1.8% of the variance and failing to visibly distinguish wine and NZ strains. Structure analysis of a selection of 100 of these strains converged with 200,000 iterations of burn in and 400,000 iterations of analysis. The resulting inferred structure at K=3 clearly distinguishes the majority of wine-group strains from the NZ population and correctly identifies the mosaic population as being highly admixed, and the EU subpopulations as more homogenous (Figure 3). The NZ samples do not form a single population but instead appear to be primarily the result of admixture between the European wine group and a second population. The Argentinian sample which clustered with NZ in the PCA is clearly distinguished as being a different population to the other wine-group samples.

Phylogenetic analysis of Ty LTRs

Phylogenetic analysis using Maximum Likelihood (ML) was used to estimate the evolutionary history of each *Ty* family present in the NZ strains. Analysis of LTRs from SGRP strains of *S. cerevisiae* and *S. paradoxus* alongside those from the NZ population allowed the identification of ancestral shared insertions (i.e. those that occurred before NZ population isolation) and insertions unique to the NZ population occurring since colonisation. No new evidence of horizontal transfer of *Ty* elements between species was recovered here. ML bootstrapping support for branches ranges from 0-100%mlBP (Maximum Likelihood Bootstrap) in each phylogeny; topologies are however plausible and consistent with previous results (Carr et al. 2012; Bleykasten-Grosshans et al. 2013). Returning poor support values is common for TE phylogenies, in particular LTR trees, regardless of inference method (e.g. Benachenhou et al., 2013).

Figure 4 displays the *Ty1/2* superfamily phylogeny. Elements of the *Ty1/2* superfamily have been active in NZ populations since colonisation (indicated by short-branched sequences), and *Ty1* activity has resulted in 18 unique solo LTRs (the result of LTR-LTR recombination) and three LTRs associated with full-length elements. Evidence of two highly active sub-lineages: *Ty1/2* hybrid and *Ty2* sub-lineage unique to the NZ strains (19 and 18 LTRs associated with internal coding regions, respectively) can also be observed (black vertical lines in Figure 4). The placement of *Ty2* within the *S. cerevisiae* *Ty1* clade is a known artifact: Carr et al. (2012) previously confirmed the origin of *Ty2* as a horizontal transfer event from *S. mikatae* into *S. cerevisiae*. *Ty2* sequences from *S. mikatae* were omitted here due to the focus on activity in NZ *S. cerevisiae*.

Supplementary Figures S2 and S3 display phylogenies for families *Ty3* and *Ty4*, respectively. The elements of both *Ty3* and *Ty4* have been minimally active since NZ colonisation; *Ty3* possesses 25 short-branched unique insertions, although none of these are associated with full-length elements. As *Ty3* is active during meiosis (Bilanchone et al., 2015), these unique insertions may be indicative of the number of mating events having occurred in the NZ population since colonisation. A further five long-branched *Ty3* insertions unique to NZ strains provide evidence of past activity of LTRs that have since accumulated mutations. *Ty4* is present in the NZ population with three unique short-branched insertions typical of relatively recent activity in the family, and four long-branched insertions displaying the accumulated mutations of aged insertions. The placement of the four long-branched NZ LTRs nested within *S. paradoxus* sequences may be the result of long-branched attraction, particularly given the poor support values for these branches (38%mlBP and 59%mlBP, respectively).

NZ strains possess only ancestral *Ty5* insertions (data not shown), therefore it is likely that the extinction of this family in *S. cerevisiae* predates the isolation of the NZ population. Trees for Ty1/2, Ty3, Ty4 and Ty5 are available as Supplementary Files S1-S4.

Differentiating features of NZ isolates

To identify candidate genes which may be under selection, Bayescan was used on a subset of 7,226 SNPs from the combined wine-NZ data, filtered to remove rare SNPs, in order to detect SNPs with differing frequencies in the wine and NZ groups. Adequate convergence was achieved with the following parameters: "Burn in: 100,000, Thinning interval: 100, Sample size: 5,000, Nb of pilot runs: 200, Length of each pilot run: 5,000". This was confirmed with Gelman and Rubin's diagnostic, which gives a by-site point scale reduction factor average of 1.0 (max 1.06), indicating the variance between and within chains is approximately equal. Overall Fst between the populations were calculated at less than 0.05, indicating low differentiation (Hartl, 1980). Using a strict False Discovery Rate of 1%, 18 candidate SNPs were identified (Figure 5). Seventeen of these SNPs were within coding regions of 9 genes. Many of these SNPs are only short distances apart within each gene, and as such unlikely represent 18 unique fixation events but also some hitchhiking events. Six of the SNPs change amino acids within three proteins (OSW7, ELP2, and CWC22) associated with outer spore assembly, spliceosomes and transcription (Table 1).

Signals of selection within NZ

Various SNP filtering levels were tested to reduce computational time and remove confounding and uninformative SNPs. The filters chosen reduced the total SNP count from 107,242 to 11,631 which were then used for analysis. The Linkage model in Structure failed to adequately converge with 10,000 iterations of burn in and 20,000 iterations of analysis and greater iteration numbers were not feasible due to the computational requirements. Under the Admixture model, convergence was achieved for K=1 through K=12 with 200,000 iterations of burn in and 400,000 iterations of analysis (Figure 6). Employing the Evanno method for calculating the most likely value of K indicated the optimal number of populations at K=2; however, the graph also included smaller peaks at K=4 and K=6, indicating that these values are also plausible (Supplementary Figure S1). K=3 and K=5 produced multimodal results,

indicating convergence to two different solutions, and as such were not analysed further. ObStruct using Structure output at K=2 and 8 sampling locations as predefined populations shows these to be strongly correlated (R^2 = 0.39, p = <0.0001) and using sampling locations with Structure output for K=4 had a weaker correlation, though still highly significant (R^2 = 0.27, p = <0.0001). Using North/South islands as predefined populations found no correlation (R^2 = 0.01, p = 0.4212).

Principal component analysis was used as an alternative method to evaluate the groups identified by Structure. After filtering for LD and missing rates, 8,333 SNPs were kept for analysis from the NZ samples. Based on ancestry inferred with Structure for K=4, 74 samples were assigned to four populations with 30 samples designated as admixed. PCA on these samples separated each of the four populations along PC1 and PC2, with admixed samples clustered in between (Figure 7b). PC1 and PC2 combined explain 9.99% of the variance and clearly separate the four populations and the admixed group. Additional PCs explained no more than 4% of the variance and were worse at distinguishing assigned populations. Assigning strains to the three populations based on the two modes found at K=3 distinguishes similarly between populations 1 and 2, separated primarily along PC1, but varies in the assignment of populations 3 and 4. Using the inferred populations for K=2 assigns 63 samples to populations, with 41 samples admixed (Figure 7a). These populations also separate along the two Principal Components but are less tightly clustered. Wine-group strains were projected onto the NZ Principal Components to compare the distinguishing features between NZ strains to the variation within the wine-group population. Here they formed a tightly clustered group, distinct from any NZ subpopulations, indicating that the wine-group strains are largely homogenous for the features which distinguish populations within NZ (Figure 7c).

Copy number variants

A total of 4.6 million of reads mapped to the ORF pangenome. The median read depth per strain ranged from 20 to 75, with a median depth from all strains of 53. Of the 7,796 ORFs in the pangenome, 6,180 were present in NZ or wine-group strains to a median copy number of at least 0.5, with 5,983 present in both populations. Of these, 1,251 were enriched in NZ and 245 in the wine-group, and 4,487 had equal median copy numbers in both populations. 78 ORFS were enriched within NZ to an average copy number at least double that of the wine-group strains, and 1.26% of the total ORFs were present in both populations (Supplementary Table S2). GOrilla identified enriched terms for Process, Function and Component (Table 2, Supplementary Table S3). In total, 18 of these ORFs were associated with an enriched GO term. The most enriched terms were associated with four sugar transporter genes from the HXT family: *HXT1, HXT4, HXT7, HXT16*. Attempts to identify ORFs with lower copy number in NZ relative to the wine-group found numerous short, unverified ORFs with no GO terms associated, along with multiple mitochondrial ORFs. The 2µ plasmid was detected in every NZ strain, with an estimated median copy number of 47. In comparison, 331 of 362 wine-group strains contained the 2µ plasmid, with a median copy number of 11.5 when present.

Discussion

Previous research on NZ *S. cerevisiae* has disagreed on the position of the NZ *S. cerevisiae* subgroup within the wine yeast clade (Cromie et al, 2013; Goddard et al., 2010; Gayevskiy et al., 2016). This study provides some evidence for the placement of the NZ *S. cerevisiae* population as a recently diverged population from the rest of the wine yeast clade (see Figure 2): they are clearly genetically similar to the established wine-group but tend to be placed more basally. However, whether the NZ strains form a single clade is less clear, as phylogenetic approaches give less clarity (Figure 1). This is not surprising given the admixed nature of the NZ population (Figure 3), and the flawed assumptions of the hierarchical clustering used which ignores this. Regardless, this study provides further support for a historic range expansion of *S. cerevisiae* that correlates with European colonists to NZ, followed

by subsequent diversification and admixture. This provides clear evidence for the range expansion of this species, but as with any other study that infers past migration from current taxa, it cannot be determined with a high degree of certainty if the genomic features distinguishing NZ and wine-group populations were present in the strain(s) which originally colonised NZ, or arose through adaptive or neutral evolutionary processes thereafter.

The population structure of *S. cerevisiae* within NZ is best explained with two subpopulations; however, PCA implies that four subpopulations may be present. This suggests that after a colonisation event, dispersal into a new habitats/ranges may have has been constrained to some extent and resulted in sub-population. Given the difficulties achieving convergence and the effect that filtering had on preliminary analysis, it is hard to determine the precise number of subpopulations. Previously, 16 populations were inferred using microsatellite data at 8 loci from 369 *S. cerevisiae* isolates using Instruct (Gao et al., 2007; Knight and Goddard, 2015). As this study includes only 104 strains, it is possible that some of the original populations were absent or underrepresented. However, given the far greater number of loci used in this analysis, 8,333 as opposed to 8, the resulting population structure will be better resolved. It has been demonstrated that SNP data is significantly better at detecting admixed populations and it is possible that some admixed populations were incorrectly identified as unique subpopulations (Gärke et al., 2012). The populations found here are structured based on sampling location, agreeing with previous studies both in NZ and in European vineyards (Capece et al., 2016). There was no evidence of the Cook Strait presenting a genetic barrier between North and South Island, as has been found in several invertebrate species (Trewick et al., 2011). The presence of the 2μ plasmid in all NZ strains, and with a higher average copy number implies more frequent sexual cycles have occurred to maintain it (Futcher et al., 1988), and this is in line with inference of an outcrossing rate between 11%-42% from 850 NZ strains (Knight and Goddard, 2015).

There appeared to be lower copy numbers of mitochondrial genes in NZ; however, this is highly related to growth conditions and not informative of genetic differences between strains (Rafelski et al., 2012).

The evolutionary history of retrotransposons typically reflects that of their host on an individual basis (Bowen et al., 2003) and as a host population (Hosid et al., 2012). While previous research has explored the activity and insertion patterns of *Ty* families in multiple populations of *S. cerevisiae* (Carr et al., 2012; Bleykasten-Grosshans et al., 2013), the work here presents the first known phylogenetic analysis of LTRs within the isolated populations of NZ. Transposition activity of families *Ty3-5* reflects that of previous research; however, the divergence and high level of activity of *Ty2* in an isolated population is yet to be reported. *Ty2* is known to be a relatively recent addition to the *S. cerevisiae* genome, likely gained by horizontal transfer from *S. mikatae* (Liti et al., 2005; Carr et al., 2012; Naseeb et al., 2017) which clearly predates the colonization of NZ and may account for *Ty2*'s relative high activity levels. Czaja et al. (2020) have recently shown that elements of the *Ty1-2* superfamily have a complex history and further research is required to understand how *Ty* elements contribute to the diversity of the *S*. *cerevisiae* genome.

Comparing ORF copy number between wine-group and NZ populations reveals a significant amount of variation. The gene ontology of the most enriched ORFs indicate that specific sugar transporter genes have increased in copy number in NZ. Expansion of *HXT7* has previously been demonstrated in experimental evolution studies under glucose limitation (Brown et al., 1998; Kao and Sherlock, 2008), along with contraction occurring under high glucose concentrations (Wenger et al., 2011). Other *HXT* genes enriched are involved in transportation of different sugars and may be expanded or contracted under high or low sugar concentrations. As such, this result correlates with observation of routes to adaptation in short term experimental studies and suggests an adaption towards altered sugar conditions in NZ compared to other wine-group strains. *S. cerevisiae* has been isolated from NZ soils

and native forests (Knight and Goddard, 2015) and one explanation is this represents an adaptation to habitats other than grape juice. These *HXT* genes represent only 4 of 78 ORFs expanded in NZ, and only 18 of those 78 were associated with enriched GO terms. Further investigation of these other enriched ORFs may reveal additional environmental adaptations in *S. cerevisiae*; however, we suggest those 60 ORFs not associated with enriched GO terms may result from neutral processes.

Overall, it is clear that the *S. cerevisiae* population in NZ has a number of genetic features differentiating it from the closely related EU population and are reasonably the result of microbial range expansion. While it is impossible to determine if these features have arisen since colonisation or were present in the *S. cerevisiae* strain(s) which founded the NZ population, some appear to have possible adaptive value. Within NZ, *S. cerevisiae* is highly geographically structured at a local level, and this structure is determined by patterns of genetic variation not found within wine-group strains. It is evident that the colonisation of NZ has produced a distinct group of *S. cerevisiae* wine-group strains, providing further support for theories of microbial range expansion emphasizing non-environmental factors. The difference in frequencies of a number of synonymous SNPs indicates that much of the genomic variation has arisen by neutral processes, and while evidence for adaptive selection may be seen in copy number variation, it is unclear if the majority of genes enriched in NZ arose from adaptive processes. Although this study has determined some of the genomic changes this colonisation has caused within *S. cerevisiae* in NZ, it has also identified a number of NZ and American strains which are differentiated from other wine-group strains by various genomic features. These may warrant further investigation to determine how these differences arose, the relationship of these American strains to the NZ population, and if the novel genetic diversity of these may contribute to ongoing work aimed at improving wine production.

Acknowledgements

The authors would like to acknowledge Martin Carr for providing *Ty* LTR datasets.

Reference List

Aa, E., Townsend, J.P., Adams, R.I., Nielsen, K.M. and Taylor, J.W. (2006) Population structure and gene evolution in saccharomyces cerevisiae. *FEMS Yeast Research,* 6(5) 702-715.

Almeida, P., Barbosa, R., Bensasson, D., Gonçalves, P. and Sampaio, J.P. (2017) Adaptive divergence in wine yeasts and their wild relatives suggests a prominent role for introgressions and rapid evolution at noncoding sites. *Molecular Ecology,* 26(7) 2167-2182.

Almeida, P., Barbosa, R., Zalar, P., Imanishi, Y., Shimizu, K., Turchetti, B., Legras, J., Serra, M., Dequin, S. and Couloux, A. (2015) A population genomics insight into the mediterranean origins of wine yeast domestication. *Molecular Ecology,* 24(21) 5412-5427.

Andrews, S. (2010) *FastQC: A Quality Control Tool for High Throughput Sequence Data.* [online]. Available from http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Benachenhou, F., Sperber, G.O., Bongcam-Rudloff, E., Andersson, G., Boeke, J.D. and Blomberg, J. (2013) Conserved structure and inferred evolutionary history of long terminal repeats (LTRs). *Mobile DNA,* 4(1) 5.

Bilanchone, V., Clemens, K., Kaake, R., Dawson, A.R., Matheos, D., Nagashima, K., Sitlani, P., Patterson, K., Chang, I., Huang, L. and Sandmeyer, S. (2015) Ty3 retrotransposon hijacks mating yeast RNA processing bodies to infect new genomes. *PLoS Genetics,* 11(9) e1005528.

Bleykasten-Grosshans, C., Friedrich, A. and Schacherer, J. (2013) Genome-wide analysis of intraspecific transposon diversity in yeast. *BMC Genomics,* 14(1) 399.

Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics,* 30(15) 2114-2120.

Bowen, N.J., Jordan, I.K., Epstein, J.A., Wood, V. and Levin, H.L. (2003) Retrotransposons and their recognition of pol II promoters: A comprehensive survey of the transposable elements from the complete genome sequence of schizosaccharomyces pombe. *Genome Research,* 13(9) 1984-1997.

Brown, C.J., Todd, K.M. and Rosenzweig, R.F. (1998) Multiple duplications of yeast hexose transport genes in response to selection in a glucose-limited environment. *Molecular Biology and Evolution,* 15(8) 931-942.

Capece, A., Granchi, L., Guerrini, S., Mangani, S., Romaniello, R., Vincenzini, M. and Romano, P. (2016) Diversity of saccharomyces cerevisiae strains isolated from two italian wine-producing regions. *Frontiers in Microbiology,* 7 1018.

Capella-Gutiérrez, S., Silla-Martínez, J.M. and Gabaldón, T. (2009) trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics,* 25(15) 1972-1973.

Carr, M., Bensasson, D. and Bergman, C.M. (2012) Evolutionary genomics of transposable elements in saccharomyces cerevisiae. *PloS One,* 7(11) e50978.

Cromie, G.A., Hyma, K.E., Ludlow, C.L., Garmendia-Torres, C., Gilbert, T.L., May, P., Huang, A.A., Dudley, A.M. and Fay, J.C. (2013) Genomic sequence diversity and population structure of saccharomyces cerevisiae assessed by RAD-seq. *G3 (Bethesda, Md.),* 3(12) 2163-2171.

Czaja, W., Bensasson, D., Ahn, H.W., Garfinkel, D.J. and Bergman, C.M. (2020) Evolution of Ty1 copy number control in yeast by horizontal transfer and recombination. *PLoS Genetics,* 16(2) e1008632.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T. and Sherry, S.T. (2011) The variant call format and VCFtools. *Bioinformatics,* 27(15) 2156-2158.

Duan, S., Han, P., Wang, Q., Liu, W., Shi, J., Li, K., Zhang, X. and Bai, F. (2018) The origin and adaptive evolution of domesticated populations of yeast from far east asia. *Nature Communications,* 9(1) 2690.

Dunn, B., Richter, C., Kvitek, D.J., Pugh, T. and Sherlock, G. (2012) Analysis of the saccharomyces cerevisiae pan-genome reveals a pool of copy number variants distributed in diverse yeast strains from differing industrial environments. *Genome Research,* 22(5) 908-924.

Eden, E., Navon, R., Steinfeld, I., Lipson, D. and Yakhini, Z. (2009) GOrilla: A tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics,* 10(1) 48.

Evanno, G., Regnaut, S. and Goudet, J. (2005) Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Molecular Ecology,* 14(8) 2611-2620.

Foissner, W. (2007) Dispersal and biogeography of protists: Recent advances. *Japanese Journal of Protozoology,* 40(1) 1-16.

Foll, M. and Gaggiotti, O. (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A bayesian perspective. *Genetics,* 180(2) 977-993.

Futcher, B., Reid, E. and Hickey, D.A. (1988) Maintenance of the 2 micron circle plasmid of saccharomyces cerevisiae by sexual transmission: An example of a selfish DNA. *Genetics,* 118(3) 411-415.

Gao, H., Williamson, S. and Bustamante, C.D. (2007) A markov chain monte carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. *Genetics,* 176(3) 1635-1651.

Garcia-Elfring, A., Barrett, R., Combs, M., Davies, T., Munshi-South, J. and Millien, V. (2017) Admixture on the northern front: Population genomics of range expansion in the white-footed

mouse (peromyscus leucopus) and secondary contact with the deer mouse (peromyscus maniculatus). *Heredity,* 119(6) 447.

Gärke, C., Ytournel, F., Bed'hom, B., Gut, I., Lathrop, M., Weigend, S. and Simianer, H. (2012) Comparison of SNPs and microsatellites for assessing the genetic structure of chicken populations. *Animal Genetics,* 43(4) 419-428.

Gayevskiy, V., Klaere, S., Knight, S. and Goddard, M.R. (2014) ObStruct: A method to objectively analyse factors driving population structure using bayesian ancestry profiles. *PLoS One,* 9(1) e85196.

Gayevskiy, V., Lee, S. and Goddard, M.R. (2016) European derived saccharomyces cerevisiae colonisation of new zealand vineyards aided by humans. *FEMS Yeast Research,* 16(7) fow091.

Goddard, M.R., Anfang, N., Tang, R., Gardner, R.C. and Jun, C. (2010) A distinct population of saccharomyces cerevisiae in new zealand: Evidence for local dispersal by insects and human-aided global dispersal in oak barrels. *Environmental Microbiology,* 12(1) 63-73.

Goddard, M.R. and Greig, D. (2015) Saccharomyces cerevisiae: A nomadic yeast with no niche? *FEMS Yeast Research,* 15(3) fov009.

González-Martínez, S.C., Ridout, K. and Pannell, J.R. (2017) Range expansion compromises adaptive evolution in an outcrossing plant. *Current Biology,* 27(16) 2544-2551. e4.

Havecker, E.R., Gao, X. and Voytas, D.F. (2004) The diversity of LTR retrotransposons. *Genome Biology,* 5(6) 225.

Hosid, E., Brodsky, L., Kalendar, R., Raskina, O. and Belyayev, A. (2012) Diversity of long terminal repeat retrotransposon genome distribution in natural populations of the wild diploid wheat aegilops speltoides. *Genetics,* 190(1) 263-274.

Jarić, I., Heger, T., Monzon, F.C., Jeschke, J.M., Kowarik, I., McConkey, K.R., Pyšek, P., Sagouis, A. and Essl, F. (2019) Crypticity in biological invasions. *Trends in Ecology & Evolution,* .

Jordan, I.K. and McDonald, J.F. (1999) Phylogenetic perspective reveals abundant Ty1/Ty2 hybrid elements in the saccharomyces cerevisiae genome. *Molecular Biology and Evolution,* 16(3) 419-422.

Kao, K.C. and Sherlock, G. (2008) Molecular characterization of clonal interference during adaptive evolution in asexual populations of saccharomyces cerevisiae. *Nature Genetics,* 40(12) 1499.

Katoh, K. and Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution,* 30(4) 772-780.

Keller, S.R., Olson, M.S., Silim, S., Schroeder, W. and Tiffin, P. (2010) Genomic diversity, population structure, and migration following rapid range expansion in the balsam poplar, populus balsamifera. *Molecular Ecology,* 19(6) 1212-1226.

Kim, J.M., Vanguri, S., Boeke, J.D., Gabriel, A. and Voytas, D.F. (1998) Transposable elements and genome organization: A comprehensive survey of retrotransposons revealed by the complete saccharomyces cerevisiae genome sequence. *Genome Research,* 8(5) 464-478.

King, M. (2003) *Penguin History of New Zealand.* Penguin UK.

Knight, S. and Goddard, M.R. (2015) Quantifying separation and similarity in a saccharomyces cerevisiae metapopulation. *The ISME Journal,* 9(2) 361.

Knight, S., Klaere, S., Fedrizzi, B. and Goddard, M.R. (2015) Regional microbial signatures positively correlate with differential wine phenotypes: Evidence for a microbial aspect to terroir. *Scientific Reports,* 5 14233.

Kopelman, N.M., Mayzel, J., Jakobsson, M., Rosenberg, N.A. and Mayrose, I. (2015) Clumpak: A program for identifying clustering modes and packaging population structure inferences across K. *Molecular Ecology Resources,* 15(5) 1179-1191.

Kvitek, D.J., Will, J.L. and Gasch, A.P. (2008) Variations in stress sensitivity and genomic expression in diverse S. cerevisiae isolates. *PLoS Genet,* 4(10) e1000223.

Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with bowtie 2. *Nature Methods,* 9(4) 357.

Legras, J., Merdinoglu, D., Cornuet, J. and Karst, F. (2007) Bread, beer and wine: Saccharomyces cerevisiae diversity reflects human history. *Molecular Ecology,* 16(10) 2091-2102.

Lesage, P. and Todeschini, A.L. (2005) Happy together: The life and times of ty retrotransposons and their hosts. *Cytogenetic and Genome Research,* 110(1-4) 70-90.

Letunic, I. and Bork, P. (2019) Interactive tree of life (iTOL) v4: Recent updates and new developments. *Nucleic Acids Research,* 47(W1) W256-W259.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics,* 25(14) 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics,* 25(16) 2078-2079.

Liti, G., Carter, D.M., Moses, A.M., Warringer, J., Parts, L., James, S.A., Davey, R.P., Roberts, I.N., Burt, A. and Koufopanou, V. (2009) Population genomics of domestic and wild yeasts. *Nature,* 458(7236) 337.

Liti, G., Peruffo, A., James, S.A., Roberts, I.N. and Louis, E.J. (2005) Inferences of evolutionary relationships from a population survey of LTR retrotransposons and telomeric associated sequences in the saccharomyces sensu stricto complex. *Yeast,* 22(3) 177-192.

Martínez, C., Cosgaya, P., Vásquez, C., Gac, S. and Ganga, A. (2007) High degree of correlation between molecular polymorphism and geographic origin of wine yeast strains. *Journal of Applied Microbiology,* 103(6) 2185-2195.

McGovern, P.E., Zhang, J., Tang, J., Zhang, Z., Hall, G.R., Moreau, R.A., Nunez, A., Butrym, E.D., Richards, M.P., Wang, C.S., Cheng, G., Zhao, Z. and Wang, C. (2004) Fermented beverages of pre- and proto-historic china. *Proceedings of the National Academy of Sciences of the United States of America,* 101(51) 17593-17598.

Meyer, K.M., Memiaghe, H., Korte, L., Kenfack, D., Alonso, A. and Bohannan, B.J. (2018) Why do microbes exhibit weak biogeographic patterns? *The ISME Journal,* 12(6) 1404.

Miller, M.A., Pfeiffer, W. and Schwartz, T. (2010) Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In: *2010 Gateway Computing Environments Workshop (GCE)* Ieee, 1-8.

Naseeb, S., James, S.A., Alsammar, H., Michaels, C.J., Gini, B., Nueno-Palop, C., Bond, C.J., McGhie, H., Roberts, I.N. and Delneri, D. (2017) Saccharomyces jurei sp. nov., isolation and genetic identification of a novel yeast species from quercus robur. *International Journal of Systematic and Evolutionary Microbiology,* 67(6) 2046-2052.

O'Malley, M.A. (2008) 'Everything is everywhere: But the environment selects': Ubiquitous distribution and ecological determinism in microbial biogeography. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences,* 39(3) 314-325.

Peter, J., De Chiara, M., Friedrich, A., Yue, J., Pflieger, D., Bergström, A., Sigwalt, A., Barre, B., Freel, K. and Llored, A. (2018) Genome evolution across 1,011 saccharomyces cerevisiae isolates. *Nature,* 556(7701) 339.

Plummer, M., Best, N., Cowles, K. and Vines, K. (2006) CODA: Convergence diagnosis and output analysis for MCMC. *R News,* 6(1) 7-11.

Power, J.F., Carere, C.R., Lee, C.K., Wakerley, G.L., Evans, D.W., Button, M., White, D., Climo, M.D., Hinze, A.M. and Morgan, X.C. (2018) Microbial biogeography of 925 geothermal springs in new zealand. *Nature Communications,* 9(1) 2876.

Pritchard, J.K., Stephens, M. and Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics,* 155(2) 945-959.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I. and Daly, M.J. (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics,* 81(3) 559-575.

Quinlan, A.R. and Hall, I.M. (2010) BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics,* 26(6) 841-842.

Rafelski, S.M., Viana, M.P., Zhang, Y., Chan, Y.H., Thorn, K.S., Yam, P., Fung, J.C., Li, H., Costa Lda, F. and Marshall, W.F. (2012) Mitochondrial network size scaling in budding yeast. *Science (New York, N.Y.),* 338(6108) 822-824.

Schuller, D., Cardoso, F., Sousa, S., Gomes, P., Gomes, A.C., Santos, M.A. and Casal, M. (2012) Genetic diversity and population structure of saccharomyces cerevisiae strains isolated from different grape varieties and winemaking regions. *PLoS One,* 7(2) e32507.

Stamatakis, A. (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics,* 30(9) 1312-1313.

Steensels, J., Gallone, B., Voordeckers, K. and Verstrepen, K.J. (2019) Domestication of industrial microbes. *Current Biology,* 29(10) R381-R393.

Swaegers, J., Mergeay, J., Van Geystelen, A., Therry, L., Larmuseau, M. and Stoks, R. (2015) Neutral and adaptive genomic signatures of rapid poleward range expansion. *Molecular Ecology,* 24(24) 6163-6176.

Thakur, M.P., van der Putten, W.H., Cobben, M.M.P., van Kleunen, M. and Geisen, S. (2019) Microbial invasions in terrestrial ecosystems. *Nature Reviews Microbiology,* 17(10) 621-631.

Trewick, S.A., Wallis, G.P. and Morgan-Richards, M. (2011) The invertebrate life of new zealand: A phylogeographic approach. *Insects,* 2(3) 297-325.

van der Gast, Christopher J (2015) Microbial biogeography: The end of the ubiquitous dispersal hypothesis? *Environmental Microbiology,* 17(3) 544-546.

Viel, A., Legras, J., Nadai, C., Carlot, M., Lombardi, A., Crespan, M., Migliaro, D., Giacomini, A. and Corich, V. (2017) The geographic distribution of saccharomyces cerevisiae isolates within three italian neighboring winemaking regions reveals strong differences in yeast abundance, genetic diversity and industrial strain dissemination. *Frontiers in Microbiology,* 8 1595.

Wang, Q., Liu, W., Liti, G., Wang, S. and Bai, F. (2012) Surprisingly diverged populations of S accharomyces cerevisiae in natural environments remote from human activity. *Molecular Ecology,* 21(22) 5404-5417.

Warringer, J., Zorgo, E., Cubillos, F.A., Zia, A., Gjuvsland, A., Simpson, J.T., Forsmark, A., Durbin, R., Omholt, S.W., Louis, E.J., Liti, G., Moses, A. and Blomberg, A. (2011) Trait variation in yeast is defined by population history. *PLoS Genetics,* 7(6) e1002111.

Wenger, J.W., Piotrowski, J., Nagarajan, S., Chiotti, K., Sherlock, G. and Rosenzweig, F. (2011) Hunger artists: Yeast adapted to carbon limitation show trade-offs under carbon sufficiency. *PLoS Genetics,* 7(8) e1002202.

Zheng, X., Levine, D., Shen, J., Gogarten, S.M., Laurie, C. and Weir, B.S. (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics,* 28(24) 3326-3328.

Zörgö, E., Gjuvsland, A., Cubillos, F.A., Louis, E.J., Liti, G., Blomberg, A., Omholt, S.W. and Warringer, J. (2012) Life history shapes trait heredity by accumulation of loss-of-function alleles in yeast. *Molecular Biology and Evolution,* 29(7) 1781-1789.

Figure 1

Phylogenetic tree of 500 *S. cerevisiae* strains produced from 16,672 SNPs using hierarchical clustering analysis showing the NZ *S. cerevisiae* wine group (blue), with previously identified wine group strains and EU subpopulations (green). The mosaic group (red) are characterized by admixture from other lineages. The neighbouring alpechin group (purple), isolated from olive mill wastewater, is included to root the tree.

Figure 2

Principal component analysis of 16,672 SNPs from NZ derived and other wine group strains. PC1 and PC2 explain 2.86% and 2.34% of the variance. NZ samples are clearly distinguished from most wine group strains along PC2; however, eight American strains (7 Californian, 1 Argentinian) cluster with the NZ samples. PC1 primarily separates out two known EU subpopulations within the wine group, while EU1 and EU2 both form tight clusters within the majority of wine group strains. Two other groups are separate from main wine group cluster but do not correspond to previously identified subgroups. A single NZ outlier, WTETETsf_C10, is indicated.

Figure 3

Structure analysis for K=3 of 5 alpechin, 5 mosaic, 40 wine-group (including 5 samples from each of the predefined EU1-4 subpopulations) and 50 NZ strains. The analysis converged with 200,000 iterations of burn in and 400,000 iterations of analysis. The EU subpopulations are highly homogenous, while mosaic and NZ populations appear admixed. In general, the wine-group is far more homogenous than the NZ population; however, a single Argentinian sample is clearly distinguished.

Figure 4

Maximum likelihood phylogeny using a nucleotide alignment of 353 *Ty1/2* LTR sequences analysed at 337 sites. Strong or maximally supported (≥70% mlBP) branches are indicated by * where space allows (Ty1_2.tre file is available in Supplementary Material). Scale bar indicates number of substitutions per site. / indicates this branch has been arbitrarily shortened for figure clarity. *Saccharomyces* Genome Resequencing Project (SGRP; Liti et al. 2009) *S. cerevisiae* strains in red allow NZ LTRs (yellow) unique to the NZ population to be identified. The tree is rooted with *S. paradoxus Ty1* (brown) sequences. The presence of unique, recently active lineages (short-branched sequences) of *Ty1/2* hybrids and *Ty2* sequences in the NZ strains is indicated by black vertical bars. *Ty1/2* hybrid LTRs are likely the result of recombination between elements of the *Ty1/2* superfamily that share high nucleotide identity.

Figure 5

Posterior odds of being under selection (PO) is plotted against F_{st} for 7226 SNPs, measured between NZ and wine group *S. cerevisiae* strains. The vertical line represents a false discovery cut-off of 1%,

SNPs to the right of this are identified as potential candidates for positive selection. Genomic information on these SNPs is provided in Table 2.

Figure 6

Structure analysis of 11,631 NZ SNPs using the admixture model with 200,000 iterations of burn in and 400,000 iterations of analysis for K=2 and K=4. The optimal number of populations was determined at K=2 using the Evanno method; however, K=4 was the next most optimal solution. Sampling location is strongly correlated with population assignment at both K=2 and K=4 ($p=<0.0001$), but North/South Island origin is not ($p=0.4212$).

Figure 7

Principal component analysis of 8,333 SNPs from NZ, showing how the first two Principal Components clearly distinguishes assigned populations from Structure for **(a)** K=2 and **(b)** K=4. Populations were assigned to strains with at least 0.8 inferred ancestry from that population. **(c)** Wine group strains form a tightly clustered group when projected on the NZ PCA.

Table 1

Genomic location of 18 SNPs identified by Bayescan as candidates for selection based on differing allele frequencies between NZ and the wine group population. The references allele is given for each SNP, along with the genomic feature containing the SNP and the identified function of this genomic feature. Coding SNPs are marked Y, non-coding SNPs are marked N.

Table 2

Gene ontology (GO) terms for genes enriched in NZ compared to the wine group when tested against a background of all identified *S. cerevisiae* genes in GOrilla. GO terms for Process, Function and Component were tested separately, and the p-values and enrichment values are presented here. For additional details of GO enrichment see Supplementary Table S3.

Table 1

Chr	Location	Ref/SNP	Genome feature	Codon change	Synonymous	Function
1	138667	A/G	XUT_1F-60	ncRNA	—	Xrn1-sensitive unstable transcript
2	372349	C/T	TIP1	K/K	Y	Major cell wall mannoprotein
5	517566	T/C	—	—	—	—
6	232998	T/C	OSW7	N/H	N	Protein involved in outer spore wall assembly
6	233004	T/C	OSW7	N/H	N	Protein involved in outer spore wall assembly
7	899840	T/A	XUT_7F-413	ncRNA	—	Xrn1-sensitive unstable transcript
7	899896	G/A	XUT_7F-413	ncRNA	—	Xrn1-sensitive unstable transcript
7	900087	C/T	ELP2	E/E	Y	Subunit of Elongator complex
7	900405	C/T	ELP2	L/L	Y	Subunit of Elongator complex
7	900645	C/T	ELP2	L/L	Y	Subunit of Elongator complex
7	900753	C/A	ELP2	E/E	Y	Subunit of Elongator complex
7	906672	G/A	ELP2	S/S	Y	Subunit of Elongator complex
7	983122	G/C	BRF1	S/F	N	TFIIIB B-related factor
7	1048402	G/A	CWC22	G/R	N	Spliceosome-associated protein
7	1048405	T/C	CWC22	S/P	N	Spliceosome-associated protein
7	1048409	G/A	CWC22	R/K	N	Spliceosome-associated protein
9	426350	T/C	Ty1 LTR	—	—	Transposable element
10	114907	T/C	JJJ2	Q/Q	Y	Protein of unknown function

Table 2

Process	P-value	Enrichment	Genes
pentose transmembrane transport	1.80E-06	95.53	HXT4, HXT7, HXT1
mannose transmembrane transport	5.65E-06	31.84	HXT16, HXT4, HXT7, HXT1
carbohydrate import across plasma membrane	7.35E-06	29.97	HXT16, HXT4, HXT7, HXT1
fructose transmembrane transport	9.39E-06	28.3	HXT16, HXT4, HXT7, HXT1
import across plasma membrane	5.12E-05	18.87	HXT16, HXT4, HXT7, HXT1
beta-alanine metabolic process	6.03E-05	127.37	ALD3, ALD2
beta-alanine biosynthetic process	6.03E-05	127.37	ALD3, ALD2
glucose import	1.02E-04	15.92	HXT16, HXT4, HXT7, HXT1
hexose transmembrane transport	1.15E-04	15.44	HXT16, HXT4, HXT7, HXT1
glucose transmembrane transport	1.15E-04	15.44	HXT16, HXT4, HXT7, HXT1
monosaccharide transmembrane transport	1.15E-04	15.44	HXT16, HXT4, HXT7, HXT1
carbohydrate transmembrane transport	1.30E-04	14.98	HXT16, HXT4, HXT7, HXT1
polyamine catabolic process	1.80E-04	84.91	ALD3, ALD2
carbohydrate transport	3.59E-04	11.58	HXT16, HXT4, HXT7, HXT1
Function			
pentose transmembrane transporter activity	1.80E-06	95.53	HXT4, HXT7, HXT1
mannose transmembrane transporter activity	4.26E-06	33.97	HXT16, HXT4, HXT7, HXT1
fructose transmembrane transporter activity	4.26E-06	33.97	HXT16, HXT4, HXT7, HXT1
glucose transmembrane transporter activity	5.65E-06	31.84	HXT16, HXT4, HXT7, HXT1
monosaccharide transmembrane transporter activity	7.35E-06	29.97	HXT16, HXT4, HXT7, HXT1
hexose transmembrane transporter activity	7.35E-06	29.97	HXT16, HXT4, HXT7, HXT1
sugar transmembrane transporter activity	9.39E-06	28.3	HXT16, HXT4, HXT7, HXT1
melatonin binding	3.68E-05	42.46	ENO1, ENO2, ADH1
hormone binding	3.68E-05	42.46	ENO1, ENO2, ADH1
carbohydrate:cation symporter activity	3.73E-05	20.38	HXT16, HXT4, HXT7, HXT1
carbohydrate:proton symporter activity	3.73E-05	20.38	HXT16, HXT4, HXT7, HXT1
carbohydrate transmembrane transporter activity	5.93E-05	18.2	HXT16, HXT4, HXT7, HXT1
amide binding	1.11E-04	10.27	ENO1, CPR1, ENO2, VTH2, ADH1
glyceraldehyde-3-phosphate dehydrogenase (NAD+) (non-phosphorylating) activity	1.80E-04	84.91	ALD3, ALD2
solute:proton symporter activity	1.82E-04	13.77	HXT16, HXT4, HXT7, HXT1
structural constituent of cell wall	1.82E-04	13.77	TIR4, TIR1, TIP1, HSP150
solute:cation symporter activity	2.47E-04	12.74	HXT16, HXT4, HXT7, HXT1
phosphopyruvate hydratase activity	3.58E-04	63.68	ENO1, ENO2
symporter activity	4.27E-04	11.08	HXT16, HXT4, HXT7, HXT1
Component			
cell periphery	2.37E-04	4.15	TIR4, PHO5, HXT16, HXT4, TIR1, HXT7, TIP1, HXT1, YFL051C
fungal-type cell wall	3.42E-04	6.26	TIR4, PHO5, PHO3, TIR1, TIP1, HSP150
phosphopyruvate hydratase complex	3.58E-04	63.68	ENO1, ENO2

external encapsulating structure	4.42E-04	5.97	TIR4, PHO5, PHO3, TIR1, TIP1, HSP150
cell wall	4.42E-04	5.97	TIR4, PHO5, PHO3, TIR1, TIP1, HSP150

Figure 1

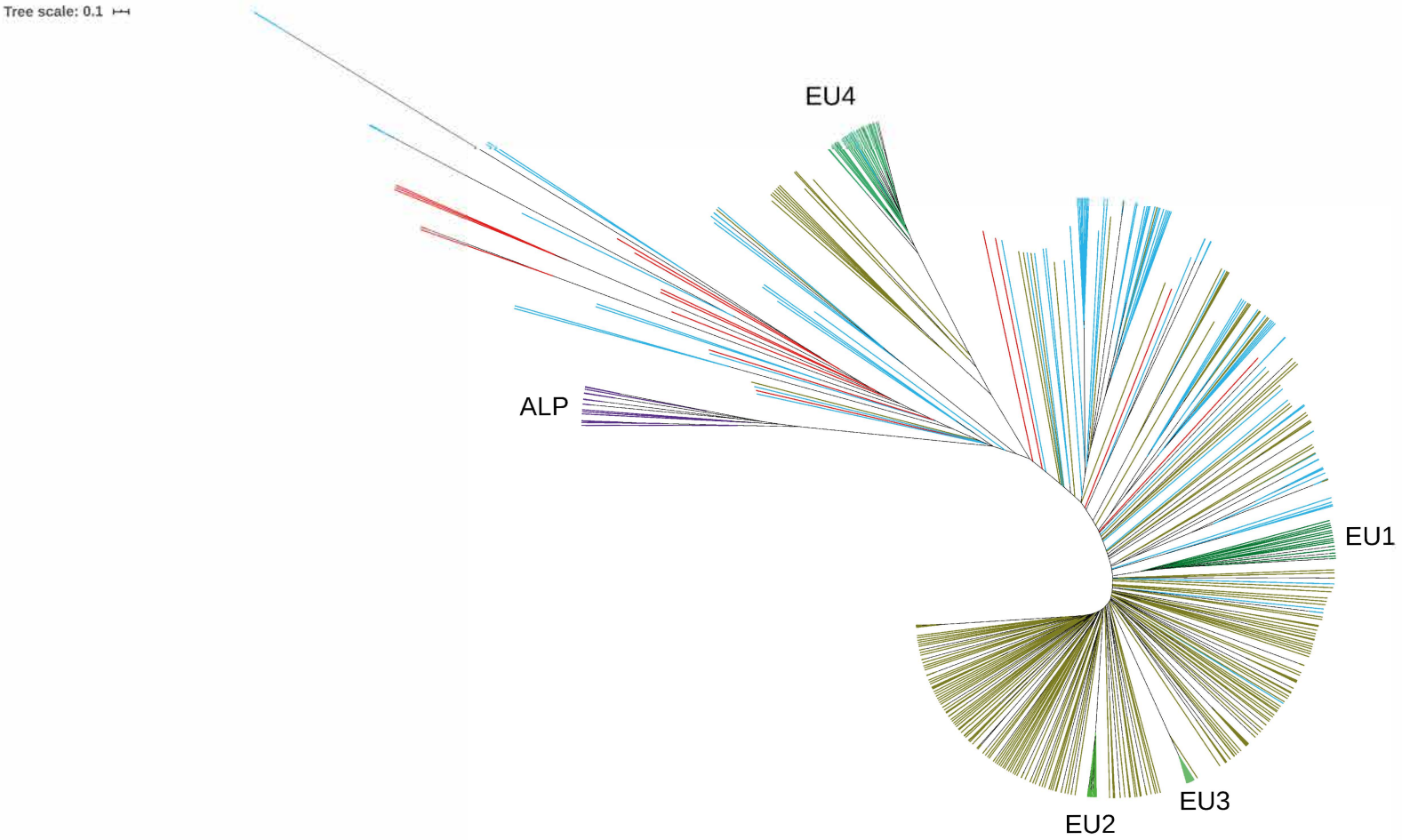


Figure 2

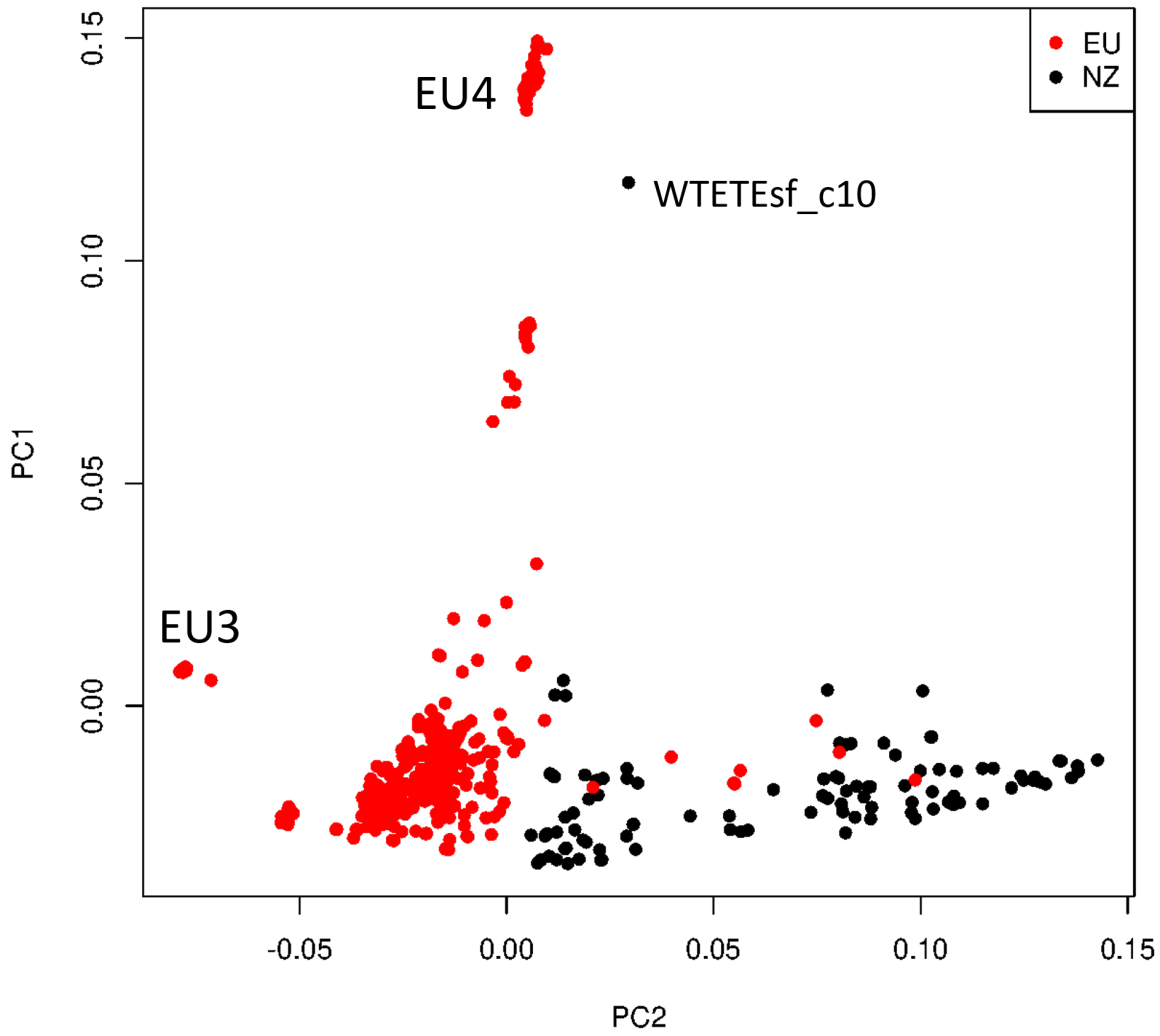

EU
NZ
EU4
WTETEsf_c10
EU3
PC1
PC2
0.15
0.10
0.05
0.00
-0.05
0.00
0.05
0.10
0.15

Figure 3

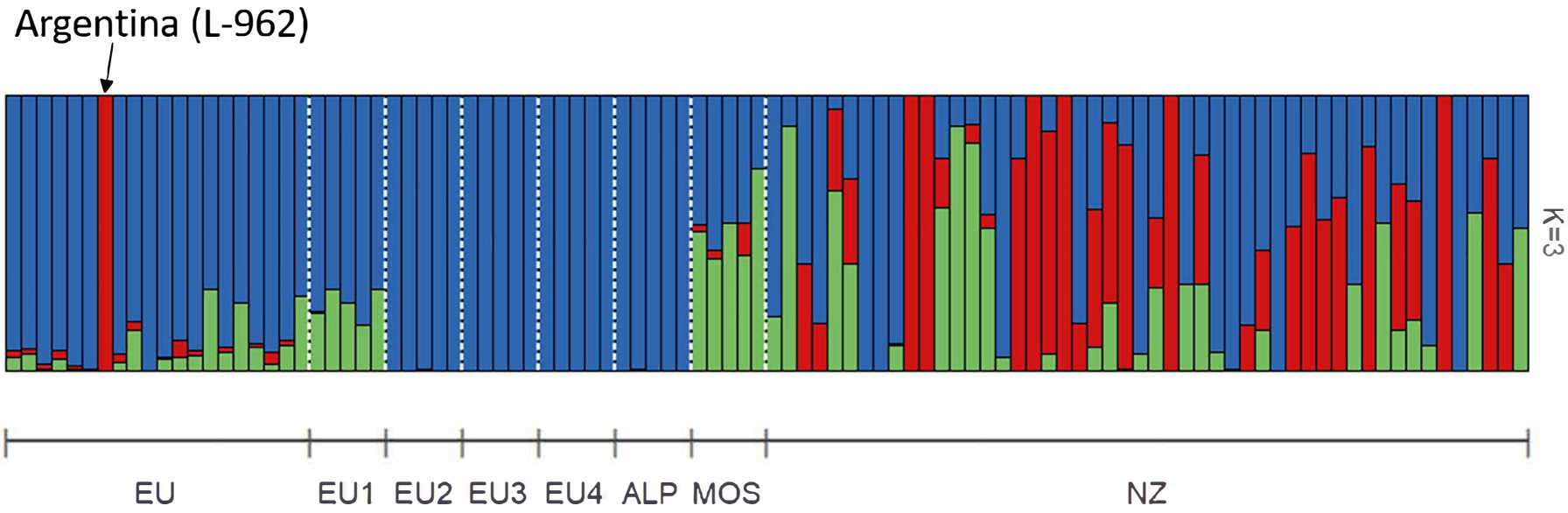


Figure 4

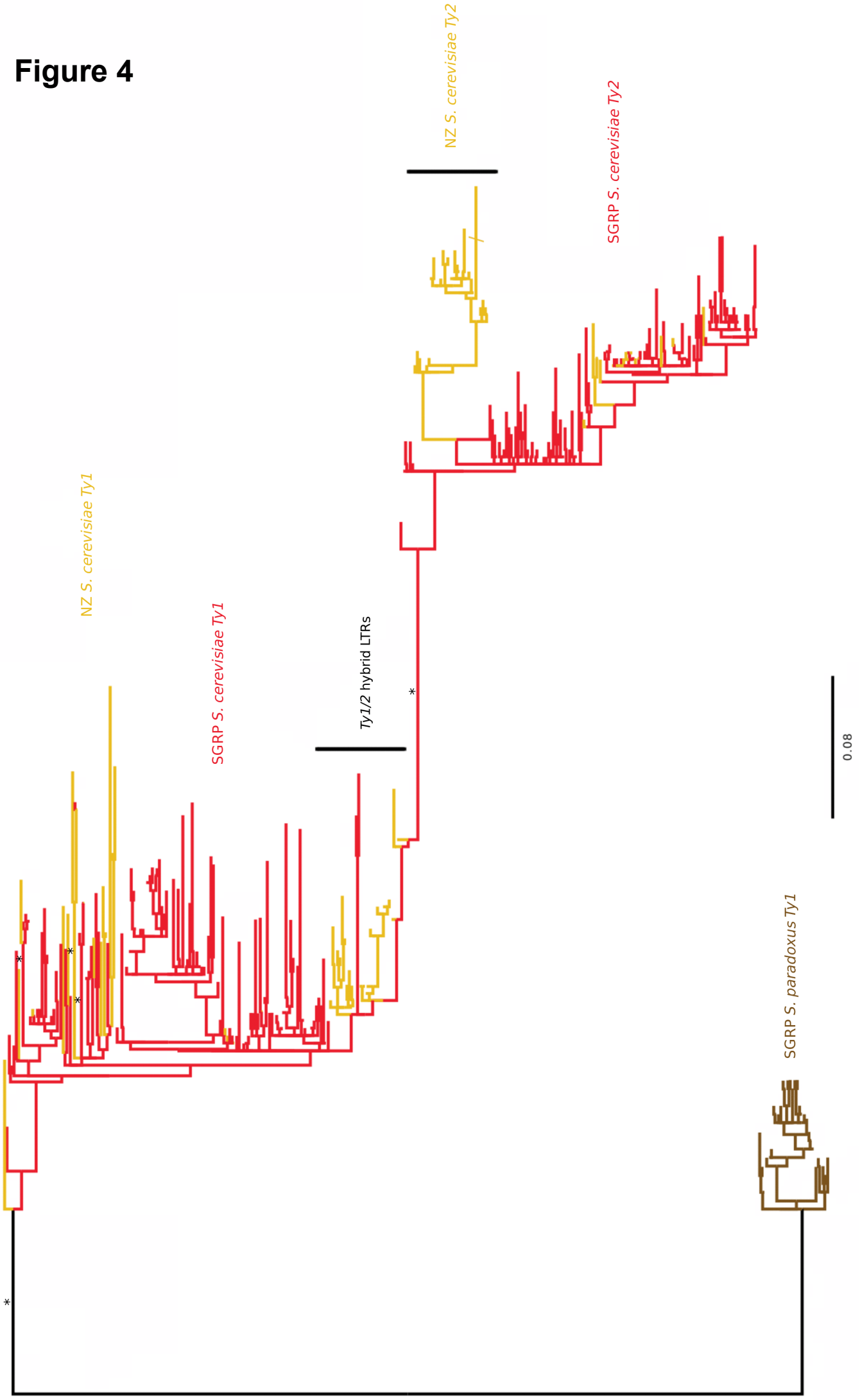

NZ S. cerevisiae Ty2
SGRP S. cerevisiae Ty2
NZ S. cerevisiae Ty1
SGRP S. cerevisiae Ty1
Ty1/2 hybrid LTRs
0.08
SGRP S. paradoxus Ty1

Figure 5

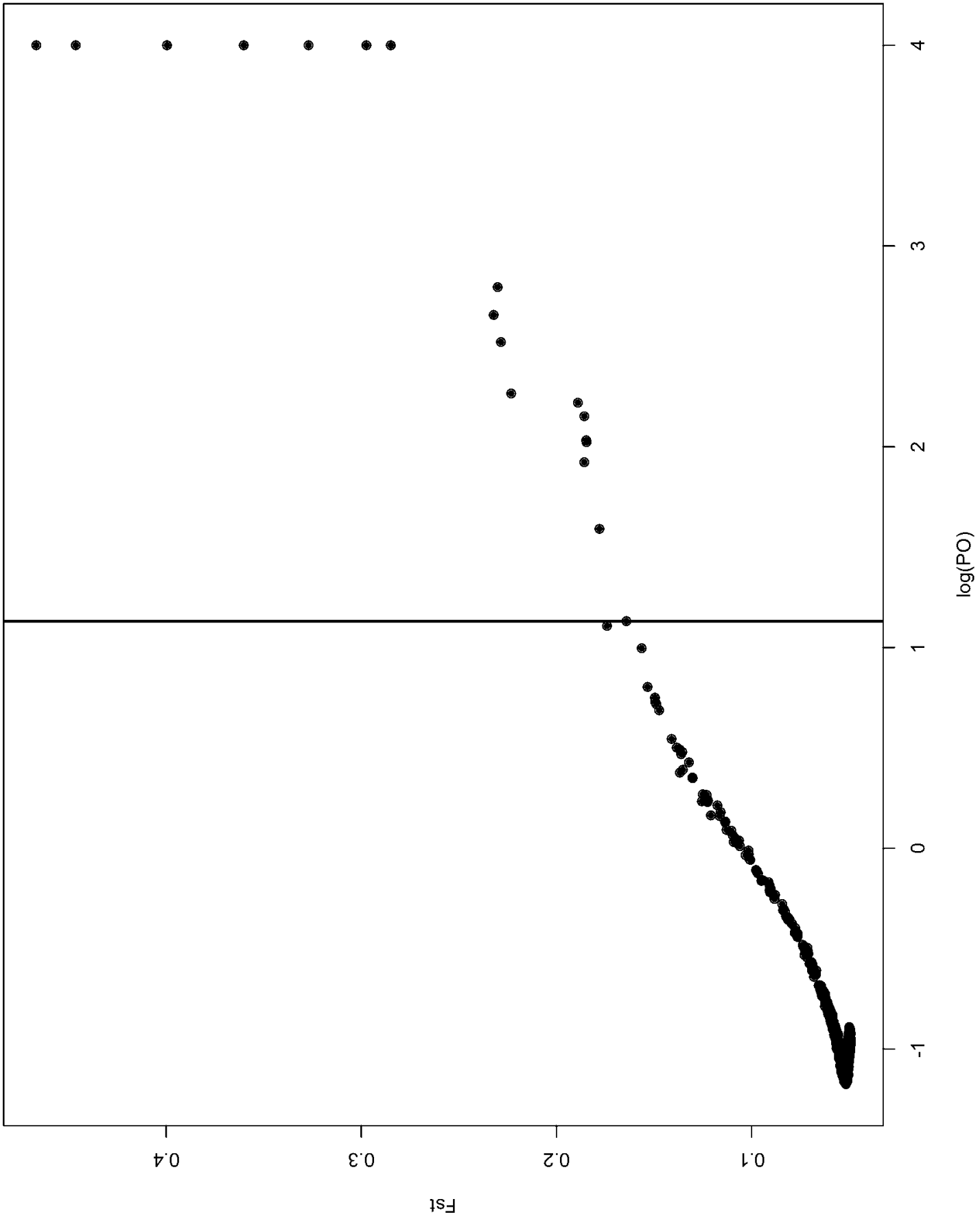


Figure 6

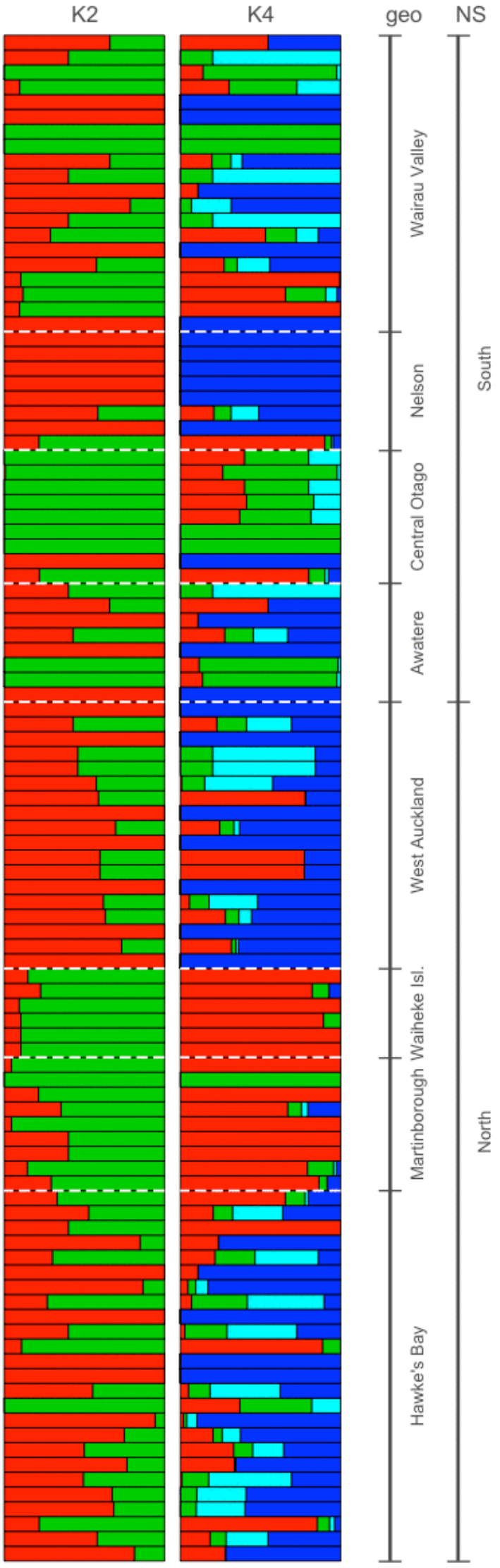
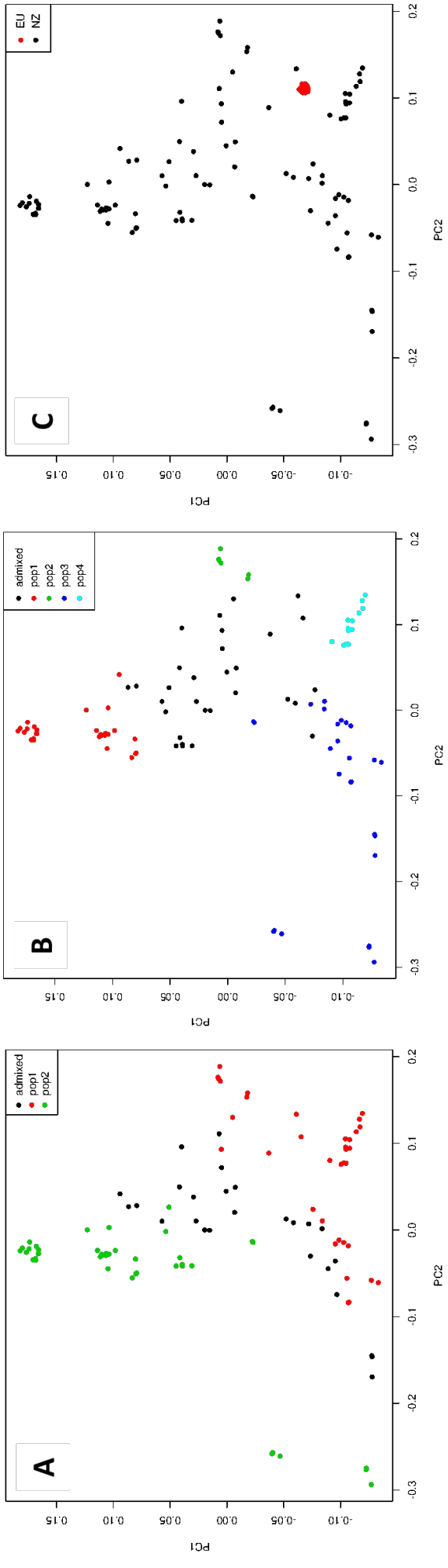

K2
K4
geo
NS
Wairau Valley
Nelson
Central Otago
Awatere
West Auckland
Waiheke Isl.
Martinborough
Hawke's Bay
South
North

Figure 7


A
admixed
pop1
pop2
B
admixed
pop1
pop2
pop3
pop4
C
EU
NZ
PC1
PC2
-0.10
-0.05
0.00
0.05
0.10
0.15
-0.3
-0.2
-0.1
0.0
0.1
0.2